Chapter XXX

Determination of Unithood and Termhood for Term Recognition

Wilson Wong
University of Western Australia, Australia

Wei Liu
University of Western Australia, Australia

Mohammed Bennamoun
University of Western Australia, Australia

ABSTRACT

As more electronic text is readily available, and more applications become knowledge intensive and ontology-enabled, term extraction, also known as automatic term recognition or terminology mining is increasingly in demand. This chapter first presents a comprehensive review of the existing techniques, discusses several issues and open problems that prevent such techniques from being practical in real-life applications, and then proposes solutions to address these issues. Keeping afresh with the recent advances in related areas such as text mining, we propose new measures for the determination of unithood, and a new scoring and ranking scheme for measuring termhood to recognise domain-specific terms. The chapter concludes with experiments to demonstrate the advantages of our new approach.

INTRODUCTION

In general, terms are considered as words used in domain-specific contexts. More specific and purposeful interpretations of terms do exist for certain applications such as ontology learning. There are two types of terms, namely, simple terms (i.e. single-word terms) and complex terms (multi-word terms). Collectively, terms constitute what is known as terminology. Terms and the tasks related to their treatments are an integral part of many applications that deal with natural language such as large-scale search engines,
Determination of Unithood and Termhood for Term Extraction

automatic thesaurus construction, machine translation and ontology learning for purposes ranging from indexing to cluster analysis. With the increasing reliance on large text sources such as the World Wide Web as input, the need to provide automated means for managing domain-specific terms arises. Such relevance and importance of terms has prompted dedicated research interests. Various names such as automatic term recognition, term extraction and terminology mining were given to encompass the tasks related to the treatments of terms.

Despite the appearance of ease in handling terms, researchers in term extraction have their share of issues due to the many ambiguities inherent to natural language and differences in the use of words. Some of the general problems include syntactic variations, structural ambiguities and inconsistent use of punctuations. The main aim in term extraction is to determine whether a word or phrase is a term which characterises the target domain. This key question can be further decomposed to reveal two critical notions in this area, namely, unithood and termhood. Unithood concerns with whether or not sequences of words should be combined to form more stable lexical units, while termhood is the extent to which these stable lexical units are relevant to some domains. Formally, unithood refers to “the degree of strength or stability of syntagmatic combinations and collocations” (Kageura and Umino, 1996), and termhood is defined as “the degree that a linguistic unit is related to domain-specific concepts” (Kageura and Umino, 1996). While the former is only relevant to complex terms, the latter concerns both simple terms and complex terms.

The research in terms and their treatments can be traced back to the field of Information Retrieval (IR) (Kageura and Umino, 1996). Many of the techniques currently available in term extraction either originated from or are inspired by advances in IR. While term extraction does share some common ground with the much larger field of IR, these two areas have some major differences. Most notably, document retrieval techniques have access to user queries to help determine the relevance of documents, while the evidences for measuring termhood are far less apparent. Based on the type of evidences employed, techniques for term extraction can be broadly categorised as linguistic-oriented and statistic-oriented. Generally, linguistic-oriented techniques rely on linguistic theories, and morphological, syntactical and dependency information obtained from natural language processing. Together with templates and patterns in the form of regular expressions, these techniques attempt to extract and identify term candidates, head-modifier relations and their context. More information is usually required to decide on the unithood and termhood of the candidates. This is when statistical techniques come into play. Evidences in the form of frequencies of occurrence and co-occurrence of the candidates, heads, modifiers and context words are employed to measure dependency, prevalence, tendency and relatedness for scoring and ranking. Consequently, it is extremely difficult for a practical term extractor or term recogniser to achieve quality results without using a combination of techniques.

As a relatively small area of research, term extraction, in recent years, enjoys a revival of interests and attentions. This is due to the increasing popularity of large-scale search engines, and the new advances in text mining techniques. Particularly, the birth of ontology learning, due to the increasing need for interoperable semantics in knowledge-based applications has necessitates a revisit to take advantage of existing advancements of automatic term recognition in order not to re-invent the wheels. A survey by Shamsfard and Barforoush (2004) has shown that many existing approaches in ontology learning merely employ isolated and non-comprehensive techniques that were not designed to address the various requirements and peculiarities in terminology. As a result, we reviewed existing automatic term recognition techniques to determine their strengths and applicability to ontology learning. We have uncovered seven issues that have effects on the performance of term extraction:
Related Content

A BPMN-Based Design and Maintenance Framework for ETL Processes

Machine Learning Techniques for Network Intrusion Detection
Tich Phuoc Tran, Pohsiang Tsai, Tony Jan and Xiangjian He (2010). *Dynamic and Advanced Data Mining for Progressing Technological Development: Innovations and Systemic Approaches* (pp. 273-299).
[www.irima-international.org/chapter/machine-learning-techniques-network-intrusion/39646/](www.irima-international.org/chapter/machine-learning-techniques-network-intrusion/39646/)

Data Visualization and Data Summary

Multi-Documents Summarization Based on TextRank and its Application in Online Argumentation Platform
Caiquan Xiong, Xuan Li, Yuan Li and Gang Liu (2018). *International Journal of Data Warehousing and Mining* (pp. 69-89).

The Use of Smart Tokens in Cleaning Integrated Warehouse Data